SPEAKER-DEPENDENT NAME DIALING IN A CAR ENVIRONMENT WITH OUT-OF-VOCABULARY REJECTION

C. S. Ramalingam, Yifan Gong, Lorin P. Netsch
Wallace W. Anderson, John J. Godfrey, and Yu-Hung Kao

Media Technologies Laboratory, Texas Instruments, Inc.,
P.O. Box 655303, MS 8374, Dallas, TX 75265, USA

ABSTRACT

In this paper we describe a system for name dialing in the car and present results under three driving conditions using real-life data. The names are enrolled in the parked car condition (engine off) and we describe two approaches for endpointing them—energy-based and recognition-based schemes—which result in word-based and phone-based models, respectively. We outline a simple algorithm to reject out-of-vocabulary names. PMC is used for noise compensation. When tested on an internally collected twenty-speaker database, for a list size of 50 and a hand-held microphone, the performance averaged over all driving conditions and speakers was 98%/92% (IV accuracy/OOV rejection); for the hands-free data, it was 98%/80%.

1. INTRODUCTION

Recently there has been a sharp increase in the use of wireless communication devices, such as mobile phones, in the car. Since such driving while dialing fits the classic definition of a "hands-busy, eyes-busy" task for which speech recognition is a natural and effective solution, not only convenience, but also safety concerns motivate the use of voice technology in this situation.

The evolving demands of speech recognition in the car and their impact on technological developments are summarized in [1]. Speaker-independent recognition in a car is addressed in [2–4] using whole word and sub-word units. Speaker-dependent recognition has been addressed by Lockwood, et al. [5, 6, and related work], whose results are based on a database containing only 4 speakers. They also do not address the issue of out-of-vocabulary (OOV) rejection. In this paper we present results that include OOV rejection using a database containing 20 speakers.

In Section 2 we describe a database collected at Texas Instruments that was used for evaluating our name dialing system. In Section 3 we outline two enrollment schemes that result in word-based and phone-based models respectively. An out-of-vocabulary rejection algorithm is described in Section 4. Recognition results are presented in Section 5 and conclusions in Section 6.

2. TEST CORPUS

The test corpus comprised speech data from 20 speakers (10 male and 10 female) collected in a car in three driving conditions: parked car (engine off), local streets within a city, and on the highway. The average speed in the city driving condition was 35 mph, but portions of the data were also collected when the car was waiting at a traffic light. For the highway condition, the speed was 60 mph or greater. The windows were closed and the radio and fan switched off. The subject was seated in the front passenger side and the data recorded simultaneously on a DAT via three typical microphones: one was placed in a cellular handset, and the other two were visor-mounted. We downsampled the speech to 8 kHz before using them in our experiments.

Data for training (parked car only) and evaluation (all conditions) were collected. Each speaker spoke 60 names (<`first-name`><`last-name`>), of which 25 were common and the remaining unique; 10 of the unique names were set aside for OOV testing, resulting in an in-vocabulary (IV) list size of 50. Additionally, there were 40 command phrases and 20 digit strings of lengths four, seven, and ten. Digit strings were not repeated, and individual digits were uniformly distributed except for zero, which occurred twice as often to allow for "oh-zero" variations, although no restrictions were placed on the speaker to pronounce one variant over the other. Each prompt sheet was read twice per session, and sessions were separated by enough time (subject to the speaker's convenience) to minimize inter-session effects. Only the name dialing portion of the corpus is of relevance to the work described in this paper.

3. ENROLLMENT

3.1. Energy-Based Endpointing

Data collected in the parked car condition were used for enrollment. Two tokens were used to build the model associated with a name. The first token was endpointed based on energy and an HMM seed model built from the extracted frames (refer to [7, chapter 6] to learn more about HMMs); the second token was used for updating the seed model. Each state in the model had a self-loop, single skip, and a progress path to the next state. Based on alignment failures during the update phase and later on recognition errors, a few problems in endpointing were corrected manually and the models retrained.

3.2. Recognition-Based Endpointing

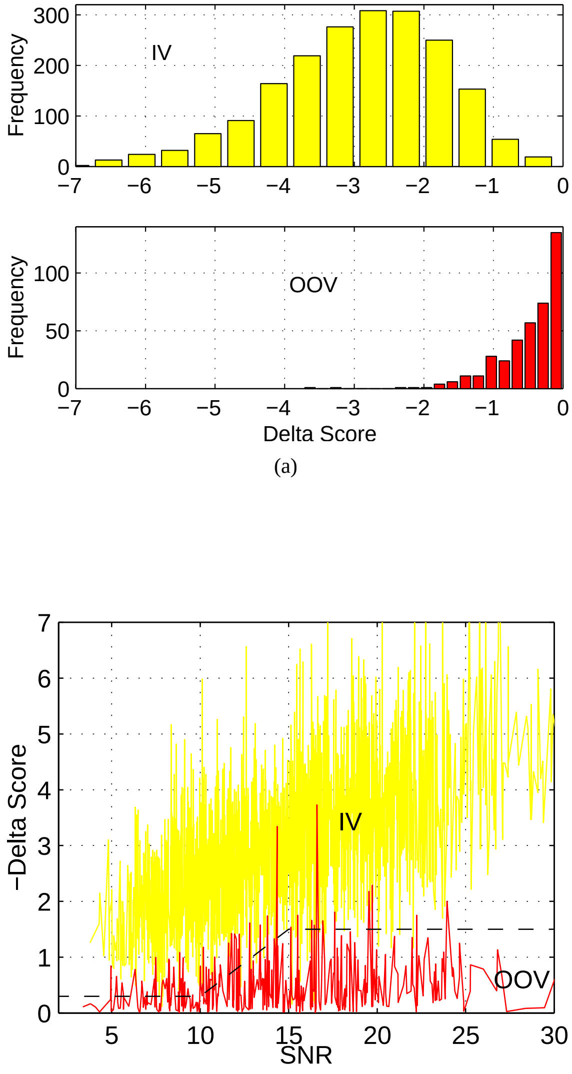
A set of mean-removed context-dependent phones trained on TIDIGITS [8] was used for endpointing the first token using simple phone-loop grammar with optional between-phone silence. After the initial alignment, each phone model was rebuilt from the enrollment utterance by assigning a state to each frame and initializing it with that frame's acoustic vector, after adding back the enrollment utterance mean. Only the mean of the distribution of each state was updated using the second token. A name-specific global variance was used since only two repetitions of each name are available. This variance was estimated from the first enrollment utterance and raised to a power; in our experiments, the exponent was 0.5. This is an ad hoc procedure whose efficacy is established only by experimental results at this point.

Compared with word models, phone-based HMMs have the advantage of benefiting from various degrees of distribution tying. Moreover, their duration can be used to obtain good OOV rejection, although in this paper we have not exploited this property.

4. OUT-OF-VOCABULARY REJECTION

In a practical system, a user is occasionally likely to utter phrases that are not on his list. A resulting substitution error, rather than a rejection, can be very costly because it will result in dialing the wrong number. Many papers in the recent past [9, 10, for example] have addressed the issue of recognition confidence in the larger context of continuous speech recognition. These methods are too complicated for our purposes. Instead, we use just the likelihood score difference between the top two hypotheses for OOV rejection.

In speaker-dependent name recognition, when an in-vocabulary item is recognized, the likelihood score is very good and the score difference between the best and second-best hypothesis is large. For an OOV item, not only is the likelihood score poorer but also the score difference between the best and next hypotheses tends to be much smaller. Therefore one can use the delta score to detect an OOV item. In Figure 1(a), histograms of the IV and OOV delta scores in the phone-based system for the city driving condition are shown, after compensating for noise using Parallel Model Combination (PMC) [11]. In Figure 1(b), the same parameter is shown as a function of signal-to-noise ratio (SNR). From that figure it is clear that delta score can be used quite reliably to distinguish between IV and OOV names, and the dashed line shows a simple SNR-dependent separation boundary.



(a)

(b)

Figure 1: (a) Histograms of IV and OOV delta scores for the cepstral front end and phone-based models in the city driving condition. PMC was used for compensating noise. (b) IV and OOV delta scores as a function of SNR.

Including the ability to reject names has the added advantage of converting most IV substitutions (a costly error) into IV rejections (a benign error), as borne out by the results presented in the next section.

5. RECOGNITION RESULTS

The names were enrolled in the parked car and tested in the following conditions: parked, city driving ("stop-and-go"), and highway driving. Results from the hand-held and one out of the two hands-free microphones are given next.

5.1. Hand-held Microphone

Only word models were used for testing the hand-held data. The analysis window was 30 ms, the frame period 20 ms, and the feature dimension 16. The filterbanks were obtained from an LPC-based spectrum [12] and a decorrelating eigen transformation was used to reduce dimensionality. A method similar to PMC was used to combat noise. The IV and OOV results are given in Table 1. These were obtained by averaging the results from the 20 speakers. For each speaker, the IV performance was tested using 100 tokens (50 names repeated twice), while the OOV performance was tested using 20 tokens (10 names repeated twice). For the parked car condition, the unused OOV names from the training session were also used for testing. The rejection boundary was adjusted so that OOV rejection was no less than 75% while maintaining about 98% IV accuracy.

Condition	IV Corr	IV Sub	OOV Rej
Parked	98.4 (100.0)	0.0	92.7
Stop-n-go	98.5 (99.9)	0.0	91.7
Highway	97.7 (99.8)	0.0	91.8
Average	98.2 (99.9)	0.0	92.2

Table 1: Results for hand-held data with word models. The numbers in parentheses represent IV correct when there is no OOV rejection.

The numbers in parentheses correspond to IV correct without OOV rejection: the recognition is almost error-free. With rejection turned on, the accuracy drops to 98.2%. OOV rejection is also very good, averaging to 92.2% over all conditions. The rejection performance was obtained by post-processing the recognition results, using delta-score information. The SNR for each file was calculated off-line and was used to determine the SNR-dependent delta threshold.

We tried experiments in which the likelihood score of the best hypothesis was also used for OOV rejection. Although it did improve rejection, the improvement was not significant. Moreover, absolute score has the disadvantage of being dependent on speaker and environment. On the other hand, delta score is far less sensitive to these variations and hence simpler to use. Similar remarks also apply to the results presented in the next section.

5.2. Hands-free Microphone

A cepstral front-end was used for testing the hands-free microphone data. The feature dimension was 16 (8 static and 8 dynamic cepstra derived from 20 mel-spaced triangular filterbanks), the window length 32 ms, and the frame period 20 ms. PMC was used to combat noise. IV and OOV results for word models are given in Table 2. Because of the simultaneous recording, the structure of the hands-free test data is identical to that described in the previous section.

Condition	IV Corr	IV Sub	OOV Rej
Parked	98.0 (99.7)	0.3	81.6
Stop-n-go	98.5 (99.7)	0.1	75.9
Highway	96.0 (99.4)	0.2	75.2
Average	97.5 (99.6)	0.2	78.6

Table 2: Results for hands-free data with word models.

As in Table 1, the numbers in parentheses represent IV accuracy without OOV rejection, which average to 99.6% over all conditions. With OOV rejection, this number drops to 97.5%. The delta threshold was adjusted so that the OOV rejection was around 75%. We did not tune the separation boundary extensively to get the best IV/OOV performance because, in the absence of data from a different microphone and/or car, one might end up tuning it to this particular test set.

Condition	IV Corr	IV Sub	OOV Rej
Parked	99.0 (99.8)	0.0	86.0
Stop-n-go	98.0 (99.8)	0.0	82.0
Highway	95.7 (99.6)	0.1	73.5
Average	97.6 (99.8)	0.0	80.5

Table 3: Results for hands-free data with phone-based models. The dashed line in Figure 1(b) shows the SNR-dependent delta threshold that was used for OOV rejection.

Results using phone models are given in Table 3. Compared with those given in Table 2, we see that phone-based models are slightly better in the stop-and-go and highway driving conditions. These models had a more restricted topology: across phone boundaries, entry was permitted only into the first state of the succeeding phone. Allowing for entry into the second state also gave slightly poorer performance.

6. CONCLUDING REMARKS

Speaker-dependent name dialing with OOV rejection capability in a noisy car environment has been shown to perform very well. We discussed two different endpointing methods that give rise to word- and phone-based models. These give very similar results, although phone models have the potential for distribution tying and improved OOV rejection based on phone duration. The OOV rejection method described in Section 4 is simple and yet at the same time quite effective. Because of database limitations, we have not tested our system using different cars or microphones. For a larger list size, OOV rejection is likely to suffer more than IV recognition, though we cannot predict the extent of the degradation. The front ends used for testing the hand-held and hands-free data, while different, are very similar in philosophy, and expected to perform very similarly. The name dialing system has been implemented on the TI C54x DSP platform.

ACKNOWLEDGMENTS

The authors wish to thank Jamil Chaoui and his team for very helpful feedback.

7. REFERENCES

[1] D. van Compernolle, "Speech Recognition in the Car: From Phone Dialing to Car Navigation," in *Proceedings of Eurospeech–97* (Rhodes, Greece), vol. V, pp. 2431–2434, Sep. 1997.

[2] D. Langmann, A. Fischer, F. Wuppermann, R. Haeb-Umbach, and T. Eisele, "Acoustic front ends for speaker-independent digit recognition in car environments," in *Proceedings of Eurospeech–97* (Rhodes, Greece), vol. V, pp. 2571–2574, Sep. 1997.

[3] R. Yang and P. Haavisto, "Noise compensation for speech recognition in car noise environments," in *Proceedings of IEEE ICASSP–95* (Detroit, MI), vol. I, pp. 433–436, May 1995.

[4] A. Fischer and V. Stahl, "Subword unit based speech recognition in car environments," in *Proceedings of IEEE ICASSP–98* (Seattle, WA), vol. I, pp. 257–260, May 1998.

[5] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view. Application to speech processing in car noise environments," *Speech Communication*, vol. 12, pp. 277–288, 1993.

[6] P. Lockwood and P. Alexandre, "Root adaptive homomorphic deconvolution schemes for speech recognition in noise," in *Proceedings of IEEE ICASSP–94* (Adelaide, Australia), vol. I, pp. 441–444, Apr. 1994.

[7] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[8] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proceedings of IEEE ICASSP–84* (San Diego, CA), vol. III, pp. 42.11.1–42.11.4, Mar. 1984.

[9] S. R. Young, "Detecting misrecognitions and out-of-vocabulary words," in *Proceedings of IEEE ICASSP-94* (Adelaide, Australia), vol. II, pp. 21–24, Apr. 1994.

[10] M. Weintraub *et al.*, "Neural network based measures of confidence for word recognition," in *Proceedings of IEEE ICASSP-97* (Munich, Germany), vol. II, pp. 887–890, Apr. 1997.

[11] M. J. F. Gales and S. J. Young, "HMM recognition in noise using parallel model combination," in *Proc. European Conf. on Speech Technology* (Berlin, Germany), vol. II, pp. 837–840, 1993.

[12] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.